

IBMEA: Exploring Variational Information Bottleneck for Multi-modal Entity Alignment

Taoyu Su
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University
of Chinese Academy of Sciences
Beijing, China
sutaoyu@iie.ac.cn

Jiawei Sheng*
Institute of Information Engineering,
Chinese Academy of Sciences
Beijing, China
shengjiawei@iie.ac.cn

Shicheng Wang
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University
of Chinese Academy of Sciences
Beijing, China
wangshicheng@iie.ac.cn

Xinghua Zhang
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University
of Chinese Academy of Sciences
Beijing, China
zhangxinghua@iie.ac.cn

Hongbo Xu
Institute of Information Engineering,
Chinese Academy of Sciences
Beijing, China
hbxu@iie.ac.cn

Tingwen Liu
Institute of Information Engineering,
Chinese Academy of Sciences
School of Cyber Security, University
of Chinese Academy of Sciences
Beijing, China
liutingwen@iie.ac.cn

Abstract

Multi-modal entity alignment (MMEA) aims to identify equivalent entities between multi-modal knowledge graphs (MMKGs), where entities can be associated with related images. Most existing studies rely heavily on the automatically learned multi-modal fusion modules, which may allow redundant information such as misleading clues in the generated entity representations, impeding the feature consistency of equivalent entities. To this end, we propose a variational framework for MMEA via information bottleneck, termed as IBMEA, by emphasizing alignment-relevant information while suppressing alignment-irrelevant information in entity representations. Specifically, we first develop multi-modal variational encoders that represent modal-specific features as probability distributions. Then, we propose four modal-specific information bottleneck regularizers to limit the misleading clues in the modal-specific entity representations. Finally, we propose a modal-hybrid information contrastive regularizer to integrate modal-specific representations and ensure the similarity of equivalent entities between MMKGs to achieve MMEA. We conduct extensive experiments on 2 cross-KG and 3 bilingual MMEA datasets. Experimental results demonstrate that our model consistently outperforms previous state-of-the-art methods, and also shows promising and robust performance especially in the low-resource and high-noise data scenarios.

CCS Concepts

• **Information systems** → **Data mining**; **Multimedia information systems**; **Information integration**;

*Corresponding author.



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0686-8/24/10
<https://doi.org/10.1145/3664647.3680954>

Keywords

Multi-modal entity alignment, Multi-modal knowledge graph, Information bottleneck

ACM Reference Format:

Taoyu Su, Jiawei Sheng, Shicheng Wang, Xinghua Zhang, Hongbo Xu, and Tingwen Liu. 2024. IBMEA: Exploring Variational Information Bottleneck for Multi-modal Entity Alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664647.3680954>

1 Introduction

Recent years have witnessed the booming of *multi-modal knowledge graphs (MMKGs)*, which extend traditional *knowledge graphs (KGs)* by introducing multi-modal data, such as the relevant images of entities, to provide physical world meanings to the symbols of KGs. Along this line, various downstream applications can be supported, such as visual question answering [52], recommendation systems [9, 48] and other applications [20, 50, 53]. However, MMKGs are usually constructed from separate multi-modal resources for different purposes, often suffering from the issue of incompleteness and limited coverage [24]. Therefore, the task of *multi-modal entity alignment (MMEA)* has been proposed, which aims to identify equivalent entities between two MMKGs, by accessing the multi-modal information of entities, such as their structures, relations, attributes, and images. In this way, one MMKG can retrieve and acquire useful knowledge from other MMKGs.

To achieve MMEA, a prominent challenge is how to exploit the feature consistency of equivalent entities between MMKGs from their diverse and abundant multi-modal information. Pioneer methods such as PoE [24] concatenate all modality features to generate holistic multi-modal entity representations, but neglect the diverse importance of different modality features. Subsequently, effective methods [3, 13, 22, 23] attempt sophisticated multi-modal fusion

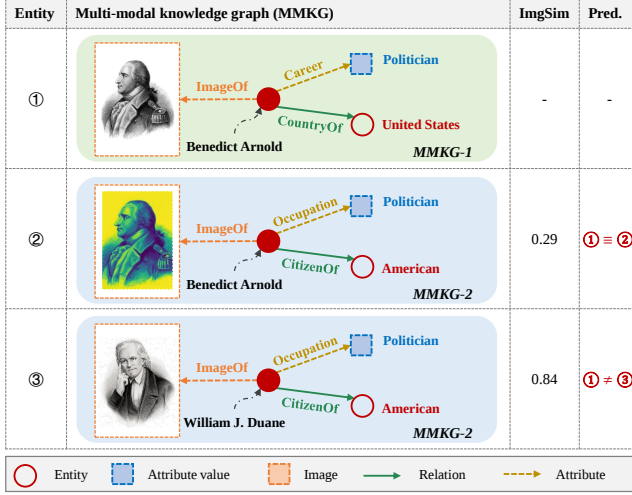


Figure 1: An example of the MMEA task between MMKGs, where *ImgSim* denotes the similarity of the images. Given *Entity_1* in *MMKG-1*, the model aims to predict *Entity_2* from candidate entities in *MMKG-2* as the true entity.

modules to focus more on the important modality features, benefiting to emphasize consistent features for equivalent entities. For example, EVA [23] learns global attention weights for different modalities, and MEAformer [3] further develops dynamic modality weights specifically for different entities. However, these methods rely heavily on the automatically learned fusion module, which can be hampered by the alignment-irrelevant information such as *misleading clues*¹ in the entities' multi-modal information, especially when there are only low-resource or high-noise training data.

In contrast, we argue that it is crucial to consider the misleading clues in modalities for MMEA. As shown in Figure 1, given *Entity_1* in *MMKG-1*, we are going to search the candidate entities (e.g., *Entity_2*, *Entity_3*, and so on) from *MMKG-2*, and identify *Entity_2* as the true equivalent entity. However, we accidentally find that the image of *Entity_1* has higher similarity² with *Entity_3* due to the similar background color. Even though, we can still identify *Entity_1* and *Entity_2* as the equivalent entities by recognizing the details of human faces. Actually, in this case, the background color reflects the misleading clue that is alignment-irrelevant for MMEA. Existing methods usually struggle³ with this case since the multi-modal fusion modules may hardly discern the alignment-relevant information from the entire image information. This inspires us to emphasize the alignment-relevant information explicitly, and simultaneously surpass the alignment-irrelevant information in modalities, to relieve the impact of misleading clues.

For this purpose, we explore *information bottleneck* (IB) [35, 36] as a potential solution. In general, the IB principle aims to learn an ideal representation that makes a trade-off between the fully data descriptive information and the task predictive information [36].

¹It is also usually called shortcuts in researches [14].

²Note that we select examples from real datasets, and obtain image features from ResNet-152 to calculate cosine similarity. For details, please see Sec. 4.4.2.

³For details, we show empirical results in Figure. 5, 6, and 7.

For the MMEA task, we expect the multi-modal encoders of different modalities to extract alignment-relevant features as the task predictive information, while appropriately ignoring certain alignment-irrelevant information in the data descriptive information. In this way, the model is required to focus more on the necessary information for MMEA from the contexts of entity pairs between MMKGs, thus limiting the redundant information in entity representations and improving the robustness to misleading clues.

Following the above idea, we propose a novel variational framework for **Multi-modal Entity Alignment via Information Bottleneck**, termed as **IBMEA**. Specifically, we first devise multi-modal variational encoders to capture the multi-modal features of each entity, including its graph structures, images, relations and attributes, and then represent them with probability distributions. Afterward, we propose two kinds of multi-modal information regularizers: 1) the modal-specific information bottleneck regularizer, which takes each modal-specific representation as input, surpassing alignment-irrelevant information and emphasizing alignment-relevant information, 2) the modal-hybrid information contrastive regularizer, which adaptively fuses all modal-specific representations to generate modal-hybrid representations, and further enhances entity similarity between entity pairs to achieve MMEA. Our major contributions can be summarized as follows:

- We introduce a novel perspective for MMEA with IB, and propose a variational framework as IBMEA. To our knowledge, we are the first to explore IB to alleviate misleading clues in MMEA.
- We propose two kinds of regularizers to refine the modal-specific and modal-hybrid features, which surpass alignment-irrelevant information and emphasize alignment-relevant information, improving the robustness with misleading clues.
- Experiments indicate that our model outperforms the comparison state-of-the-art methods on 5 benchmarks, and obtains promising and robust results in low-resource and high-noise scenarios.

2 Preliminaries

2.1 Task Formulation

Formally, the **multi-modal knowledge graph (MMKG)** can be defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V})$, where $\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}$ are the sets of entities, relations, attributes and visual images, respectively. Therefore, the triples are defined as $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where each entity $e \in \mathcal{E}$ can be linked to attributes and images. Building on previous research [3, 22, 23], we mainly focus four modalities $m \in \{g, v, r, a\}$ of entities, including graph structures g , visual images v , neighboring relations r and attributes a .

Based upon, **multi-modal entity alignment (MMEA)** aims to identify equivalent entities from two different MMKGs. Formally, given two MMKGs $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ with their relational triples and multi-modal attributes, the goal of MMEA task is to identify equivalent entity pairs $\{(e^{(1)}, e^{(2)}) | e^{(1)} \in \mathcal{G}^{(1)}, e^{(2)} \in \mathcal{G}^{(2)}, e^{(1)} \equiv e^{(2)}\}$. During training, the model utilizes a set of pre-aligned entity pairs $\mathcal{S}$ as alignment seeds, while testing focuses on predicting equivalent entity pairs across MMKGs. In this paper, we aim to achieve MMEA by learning entity representations from both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, emphasizing alignment-relevant information while surpassing alignment-irrelevant information to bridge between MMKGs.

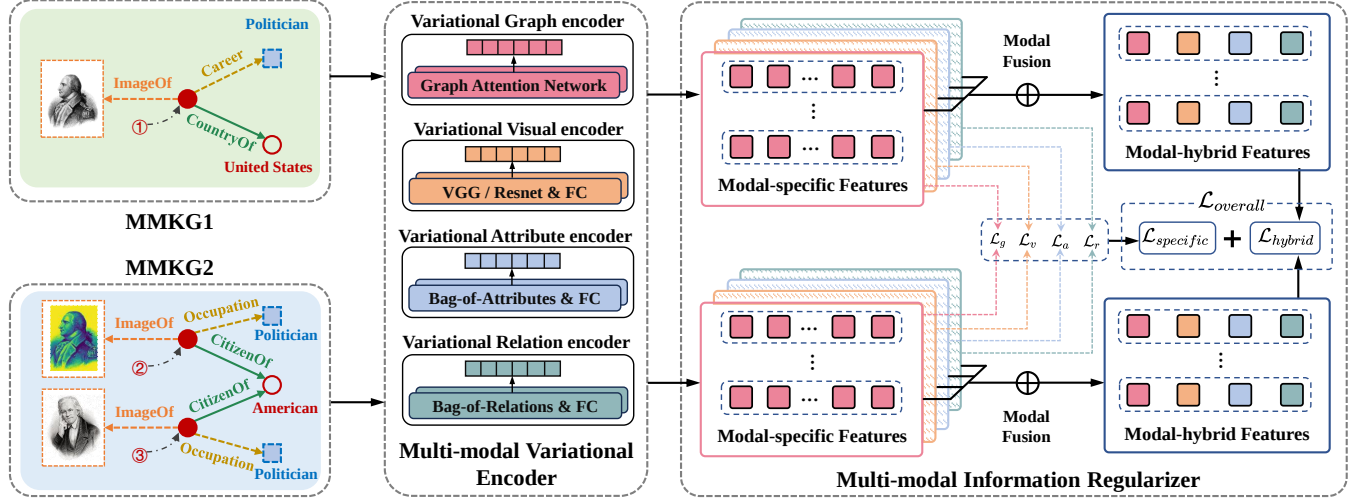


Figure 2: The framework of the proposed IBMEA for the multi-modal entity alignment task.

2.2 Information Bottleneck

We adopt the information bottleneck (IB) principle [35, 36] to learn general entity representations across MMKGs. The IB principle aims to find a maximally compressed representation Z of the input X while preserving information about the target Y , which achieved by minimizing the objective function:

$$\mathcal{L}_{IB} = \beta I(Z; X) - I(Z; Y), \quad (1)$$

where $I(\cdot; \cdot)$ denotes mutual information, and $\beta > 0$ is a Lagrangian multiplier controlling the trade-off between compression and information preservation. Minimizing the first term penalizes the information between Z and X , encouraging the latent variable Z to forget irrelevant information. Maximizing the second term encourages Z to be predictive of Y . Consequently, the IB principle enables Z to capture relevant factors for prediction while compressing irrelevant parts [6]. This aligns with the notion of Z acting as a minimal sufficient statistic of X for predicting Y [5].

Although the IB principle is appealing, computing mutual information poses computational challenges. To overcome this, the VIB (Variational Information Bottleneck) [5] method proposes a variational approximation approach for the IB objective. It minimizes the following formulation:

$$\mathcal{L}_{VIB} = \beta \mathbb{E}_{\mathbf{x}} [\text{KL} [p_{\theta}(\mathbf{z}|\mathbf{x}) || r(\mathbf{z})]] - \mathbb{E}_{\mathbf{z} \sim p_{\theta}(\mathbf{z}|\mathbf{x})} [\log q_{\phi}(\mathbf{y}|\mathbf{z})], \quad (2)$$

where $p_{\theta}(\mathbf{z}|\mathbf{x})$ is an estimate of posterior probability to generate $\mathbf{z}$ from data $\mathbf{x}$, $r(\mathbf{z})$ is an estimate of the prior probability $p(\mathbf{z})$ to describe $\mathbf{z}$, and $q_{\phi}(\mathbf{y}|\mathbf{z})$ is a parametric approximation of $p(\mathbf{y}|\mathbf{z})$ to predict target $\mathbf{y}$. Intuitively, the encoder generates $p_{\theta}(\mathbf{z}|\mathbf{x})$ with parameter θ , and the decoder achieves $q_{\phi}(\mathbf{y}|\mathbf{z})$ with parameter ϕ . Actually, Eq. (2) resembles a likelihood-prior trade-off, since the first term applies KL-divergence with uninformative prior, penalizing the too complex encoded representations, and the second term approximates the log-likelihood expectation, requiring the representations to achieve task prediction.

3 Methodology

The overall framework of IBMEA is illustrated in Figure 2. Our model mainly consists of two components: 1) *Multi-modal Variational Encoder*, which represents multi-modal information of entities from both MMKGs into random variables, and 2) *Multi-modal Information Regularizer*, which imposes IB regularization on each modal-specific representations, and facilitates MMEA with contrastive regularization upon modal-hybrid representations.

3.1 Multi-modal Variational Encoder

To capture entity features from all modalities, including the graph structure g , visual image v , relation r and attribute a . We devise the modal-specific variational encoder and generate the modal-specific feature variables as $\mathbf{z}_m$ with $m \in \{g, v, r, a\}$, respectively.

3.1.1 Variational Graph Encoder. To fully leverage the graph structures of MMKGs, we devise a variational graph encoder. Considering the neighbors of an entity can have different importance for entity alignment, we adopt (two attention heads and two layers) graph attention network (GAT) [40] to capture graph structures. To better depict the information of the graph structural feature, we represent it with a variable $\mathbf{z}_g$ of the probability distribution. Practically, we assume the variable as Gaussian distributions as widely used in variational studies [5, 16, 44], and represent $\mathbf{z}_g$ with mean vector μ_g and variance vector σ_g , derived by separate GATs with different learnable parameters [16]:

$$\begin{aligned} \mu_g &= \text{GAT}(\mathbf{A}, \mathbf{x}_g; \theta_{g,\mu}), \\ \sigma_g &= \text{GAT}(\mathbf{A}, \mathbf{x}_g; \theta_{g,\sigma}), \\ \mathbf{z}_g &\sim \mathcal{N}(\mu_g, \text{diag}(\sigma_g^2)), \end{aligned} \quad (3)$$

where $\mathbf{x}_g \in \mathbb{R}^{d_g}$ is the randomly initialized node features, $\mathbf{A}$ denotes the adjacent matrix of the given MMKG. To alleviate the negative impacts of structural heterogeneity, we encode both MMKGs with shared GATs to project entities into the same embedding space [18]. The learnable parameters can be summarized as $\theta_g \triangleq \{\theta_{g,\mu}, \theta_{g,\sigma}\}$.

3.1.2 Variational Visual, Attribute and Relation Encoder. To highlight the information of relations, attributes, and images, we devise separate fully connected layers as multi-modal encoders to learn interim representations for each modality $m' \in \{v, a, r\}$ as follows:

$$\hat{\mathbf{h}}_{m'} = \delta(\text{FC}_{m'}(\mathbf{W}_{m'}, \mathbf{x}_{m'})), \quad (4)$$

where $\mathbf{W}_{m'} \in \mathbb{R}^{d_{m'} \times d}$ and δ is the ReLU activation. Here, $\mathbf{x}_{m'} \in \mathbb{R}^{d_{m'}}$ denotes the initialized feature for m' -modality. For images, we utilize pre-trained visual encoders to obtain effective image features. For entities without images, we average the images of their neighbors as the initial feature [19]. For attributes and relations, the bag-of-attributes and bag-of-relations features [49] are employed.

Based on these interim representations, we employ multi-layer perceptrons (MLPs) to derive modal-specific feature variables. Assuming these variables follow a Gaussian distribution and employ separate MLPs to learn mean vector $\boldsymbol{\mu}_{m'}$ and variance vector $\boldsymbol{\sigma}_{m'}$:

$$\begin{aligned} \boldsymbol{\mu}_{m'} &= \text{MLP}(\hat{\mathbf{h}}_{m'}; \theta_{m', \mu}), \\ \boldsymbol{\sigma}_{m'} &= \text{MLP}(\hat{\mathbf{h}}_{m'}; \theta_{m', \sigma}), \\ \mathbf{z}_{m'} &\sim \mathcal{N}(\boldsymbol{\mu}_{m'}, \text{diag}(\boldsymbol{\sigma}_{m'}^2)), \end{aligned} \quad (5)$$

where the learnable parameters for m' -modality can be summarized as $\theta_{m'} \triangleq \{\theta_{m', \mu}, \theta_{m', \sigma}, \mathbf{W}_{m'}\}$ with $m' \in \{v, a, r\}$.

3.1.3 Multi-modal Representation Implementation. Since the modal-specific variables are intractable for neural networks, we leverage the reparameterization trick [16] to sample deterministic representations from probability distributions. For each modality $m \in \{g, v, a, r\}$, the sampled representation $\mathbf{z}_m$ is obtained as follows:

$$\mathbf{z}_m = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \text{diag}(\mathbf{I})), \quad (6)$$

where $\boldsymbol{\mu}_m$ and $\boldsymbol{\sigma}_m$ are the corresponding mean and standard deviation of m -modality, respectively. $\boldsymbol{\epsilon}$ is a standard Gaussian noise, and $\odot$ denotes the element-wise production. In this way, we obtain the deterministic modal-specific representations for entities, which can be further integrated to acquire entire (modal-hybrid) entity representations to make predictions for MMEA.

3.2 Multi-modal Information Regularizer

In this section, we introduce two kinds of regularizers from mutual information perspective, which refine the modal-specific representations with limited alignment-irrelevant information, and derive better modal-hybrid representations for MMEA.

3.2.1 Modal-specific Information Bottleneck Regularizer. As claimed in Sec. 1, there may exist alignment-irrelevant misleading clues in distinct modalities. For each modality $m \in \{g, v, a, r\}$, we expect the learned modal-specific feature variables $\mathbf{z}_m$ to contain less alignment-irrelevant information while retaining more alignment-relevant information between MMKGs. Following this idea, we propose the information bottleneck regularizer $\mathcal{L}_m$ for m -modality feature variables of the paired entities from the two MMKGs:

$$\begin{aligned} \mathcal{L}_m &= \sum_{i \in \{1, 2\}} [\beta_m I(\mathbf{Z}_m^{(i)}; \mathbf{X}_m^{(i)}) - I(\mathbf{Z}_m^{(i)}; \mathbf{Y})], \\ &= \underbrace{\beta_m [I(\mathbf{Z}_m^{(1)}; \mathbf{X}_m^{(1)}) + I(\mathbf{Z}_m^{(2)}; \mathbf{X}_m^{(2)})]}_{\text{Minimality}} - \underbrace{I(\mathbf{Z}_m^{(1)}, \mathbf{Z}_m^{(2)}; \mathbf{Y})}_{\text{Alignment}}, \end{aligned} \quad (7)$$

where $\mathbf{X}_m^{(1)}$ and $\mathbf{X}_m^{(2)}$ denotes the original m -modality features from $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$, respectively. $\mathbf{Z}_m^{(1)}$ and $\mathbf{Z}_m^{(2)}$ are obtained by Eq. (6) over the two MMKGs. $\mathbf{Y}$ denotes the supervised information, indicating whether the paired entities are equivalent. Note that *the minimality term penalizes the information from its own MMKG, while the alignment term encourages the representations to be capable of entity alignment*. In this way, the redundant information is surpassed by minimality term, alleviating alignment-irrelevant misleading clues in the modal-specific representations. On the other hand, the alignment-relevant information is simultaneously emphasized, since we require the modal to make prediction by the alignment term with tailored information. For tractable objective function, please refer to Sec. 3.3.1.

To refine modal-specific representations of all modalities, the overall IB regularizer $\mathcal{L}_{\text{specific}}$ is defined as follows:

$$\mathcal{L}_{\text{specific}} = \sum_{m \in \{g, v, a, r\}} \mathcal{L}_m. \quad (8)$$

In this way, we encourage each modality of information to remain consistency between MMKGs, benefiting to entire representations.

3.2.2 Modal-hybrid Information Contrastive Regularizer. To fully leverage information of all modalities for MMEA, we propose a modal-hybrid information contrastive regularizer. We first generate modal-hybrid feature variables by integrating all the modal-specific feature variables. Specifically, for each entity, we measure the distinct importance of its modality information with attention mechanism, and employ the attention weights to integrate modal-specific feature variables (sampled from Eq. (6)) as follows:

$$\begin{aligned} s_m &= \mathbf{q}_o^T \tanh(\mathbf{W}_m \mathbf{z}_m + b_m), \\ \alpha_m &= \frac{\exp(s_m)}{\sum_{k \in \{g, v, a, r\}} \exp(s_k)}, \\ \mathbf{z}_o &= \sum_{m \in \{g, v, a, r\}} \alpha_m \mathbf{z}_m, \end{aligned} \quad (9)$$

where α_m is the attention weight for modality m , taking the different nature of entities into consideration. In addition, $\mathbf{q}_o \in \mathbb{R}^d$, $\mathbf{W}_m \in \mathbb{R}^{d \times d_m}$ and $b_m \in \mathbb{R}^d$ are learnable parameters. In this way, we obtain modal-hybrid feature variables considering the distinct modality importance of the entity and leverage the IB-refined modal-specific feature variables.

Based on the derived modal-hybrid feature variable, we aim to measure the mutual information between paired entities from the two MMKGs. Our main intuition is that, *given an entity in the one MMKG, the equivalent entity in the other MMKG would have higher mutual information than other candidate entities*. Following this intuition, we propose the modal-hybrid information contrastive regularizer $\mathcal{L}_{\text{hybrid}}$ as follows:

$$\mathcal{L}_{\text{hybrid}} = - \underbrace{I(\mathbf{Z}_o^{(1)}; \mathbf{Z}_o^{(2)})}_{\text{Contrastive term}}. \quad (10)$$

In this way, we measure the consistency between paired entities from the two MMKGs, which fully leverages all the IB-refined modal-specific features to constitute the entire entity representations. For tractable objective function, please refer to Sec. 3.3.2.

3.3 Tractable Optimization Objective

Theoretically, direct optimization of mutual information can be intractable considering the intractable integrals [6, 29]. Therefore, we derive tractable objectives with variational approximation [38].

3.3.1 Tractable Information Bottleneck Objective. To optimize the information bottleneck regularizer in Eq. (7), we derive the variational lower bound for the minimality and alignment term.

For the **minimality term**, we measure it by Kullback-Leibler (KL) divergence [4] with variational approximation posterior distributions. Specifically, recall that we are given the original feature $\mathbf{x}_m$ of m -modality from $\mathcal{G}$ ($m \in \{g, v, r, a\}$), and we omit the subscript of different MMKGs), and the feature variable $\mathbf{z}_m$ can be approximated by the m -th modality variational encoder. We can derive the approximation upper bound [29] of the minimality term, and minimize it as the tractable objective function:

$$\begin{aligned} I(\mathbf{Z}_m; \mathbf{X}_m) &\leq \mathbb{D}_{KL}(p_\theta(\mathbf{z}_m | \mathbf{x}_m) || r(\mathbf{z}_m)) \\ &= \mathbb{D}_{KL}(\mathcal{N}(\boldsymbol{\mu}_m, \text{diag}(\boldsymbol{\sigma}_m^2)) || \mathcal{N}(0, \text{diag}(\mathbf{I}))). \end{aligned} \quad (11)$$

Here, to approximate true posterior distributions $p(\mathbf{z}_m | \mathbf{x}_m)$, we adopt the m -modality variational encoder parameterized by θ (omit subscript m of θ_m) to generate approximated posterior distribution as $p_\theta(\mathbf{z}_m | \mathbf{x}_m)$. Besides, following existing variational methods [5, 16, 44], we also assume the prior distribution $r(\mathbf{z}_m)$ as a standard Gaussian distribution $\mathcal{N}(0, \text{diag}(\mathbf{I}))$ for convenience. Such a procedure holds for both MMKG $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$. In this way, we compress redundant information in the m -modality representation.

For the **alignment term**, we also adopt variational approximation posterior distribution to render. The goal of the alignment term is to make prediction with the learned feature variables. Therefore, we rewrite the alignment term with likelihood of the entity alignment task, which is the variational lower bound for maximization as the tractable objective function:

$$\begin{aligned} I(\mathbf{Z}_m^{(1)}, \mathbf{Z}_m^{(2)}; \mathbf{Y}) &\geq \mathbb{E}_{\mathbf{z}_m^{(1)}, \mathbf{z}_m^{(2)} \sim p_\theta(\mathbf{z}_m^{(1)} | \mathbf{x}_m^{(1)}), p_\theta(\mathbf{z}_m^{(2)} | \mathbf{x}_m^{(2)})} [\log q_\phi(\mathbf{y} | \mathbf{z}_m^{(1)}, \mathbf{z}_m^{(2)})], \end{aligned} \quad (12)$$

where the variable $\mathbf{y}$ denotes whether the entity pair is equivalent, and $q_\phi(\mathbf{y} | \mathbf{z}_m^{(1)}, \mathbf{z}_m^{(2)})$ is the entity alignment decoder to make prediction. Generally, the likelihood can be achieved by binary cross-entropy, and here we use InfoNCE [39] to better consider all the candidate entities in ranking, which can be formed as:

$$\begin{aligned} I(\mathbf{Z}_m^{(1)}, \mathbf{Z}_m^{(2)}; \mathbf{Y}) &\geq \sum_{(\mathbf{z}_m^{(1)}, \hat{\mathbf{z}}_m^{(2)}) \in \mathcal{S}} [\log \exp(\cos(\mathbf{z}_m^{(1)}, \mathbf{z}_m^{(2)})/\tau) \\ &\quad - \sum_{(\mathbf{z}_m^{(1)}, \hat{\mathbf{z}}_m^{(2)}) \notin \mathcal{S}} \log \exp(\cos(\mathbf{z}_m^{(1)}, \hat{\mathbf{z}}_m^{(2)})/\tau)], \end{aligned} \quad (13)$$

where $\mathcal{S}$ is the seed alignments (i.e., pre-aligned entity pairs), τ is the temperature factor. $\hat{\mathbf{z}}_m^{(2)}$ is the negative entity obtained by randomly replacing the true entity of seed alignments.

3.3.2 Tractable Information Contrastive Objective. To optimize the contrastive term, we follow the idea of InfoMax [41] by measuring the mutual information with neural networks. Specifically, we use another discriminator D to measure the consistency between entities, and the lower bound of the contrastive term can be derived

as:

$$\begin{aligned} I(\mathbf{Z}_o^{(1)}; \mathbf{Z}_o^{(2)}) &= \mathbb{E}_{p(\mathbf{z}_o^{(1)} | \mathbf{x}^{(1)}) p(\mathbf{z}_o^{(2)} | \mathbf{x}^{(2)})} [\log D(\mathbf{z}_o^{(1)}, \mathbf{z}_o^{(2)})] \\ &\geq \sum_{(\mathbf{z}_o^{(1)}, \mathbf{z}_o^{(2)}) \in \mathcal{S}} \log(D(\mathbf{z}_o^{(1)}, \mathbf{z}_o^{(2)})) + \sum_{(\mathbf{z}_o^{(1)}, \mathbf{z}_o^{(2)}) \notin \mathcal{S}} \log(1 - D(\mathbf{z}_o^{(1)}, \mathbf{z}_o^{(2)})), \end{aligned} \quad (14)$$

where $\mathbf{z}_o^{(1)}$ and $\mathbf{z}_o^{(2)}$ are modal-hybrid variables for $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ from Eq. (9) respectively, which encode all the original modality features ($\mathbf{x} \triangleq \{x_g, x_v, x_a, x_r\}$) in the corresponding MMKG. Here we can achieve the discriminator D with inner production followed by sigmoid. Inspired by Boudiaf et al. [8], we adopt multi-similarity loss [43] to further consider the impact of different entities in contrastive learning. In this way, we encourage the modal-hybrid representations of the equivalent entities to be relevant and thereupon achieve the MMEA task with all modality information.

3.3.3 Overall Objective Function. Since we have imposed constraints on both the modal-specific and modal-hybrid representations, the overall objective function is defined in a joint paradigm as follows:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{specific}} + \mathcal{L}_{\text{hybrid}}. \quad (15)$$

In summary, the proposed IB regularizers enable us to suppress alignment-irrelevant information and emphasize the alignment-relevant information to fulfill the MMEA task.

4 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness. More analyses are detailed in **Appendix**⁴.

4.1 Experimental Settings

4.1.1 Datasets and Evaluation Metrics. In this study, we evaluate our proposed IBMEA on five multi-modal EA datasets, categorized into two types. (1) Cross-KG [24] datasets, comprising FB15K-DB15K and FB15K-YAGO15K, with 128,486 and 11,199 labeled pre-aligned entity pairs, respectively. (2) Bilingual datasets also named DBP15K [23, 32], which are including DBP15K_{ZH-EN}, DBP15K_{JA-EN} and DBP15K_{FR-EN} from the multilingual versions of DBpedia, each with approximately 400K triples and 15K pre-aligned entity pairs. EVA [23] provided images of entities for the DBP15K dataset. Following previous works [3, 22, 23], we utilize 20%, 50%, 80% of true entity pairs as alignment seeds for training on cross-KG datasets, and 30% for bilingual datasets. We use Hits@1 (H@1), Hits@10 (H@10), and Mean Reciprocal Rank (MRR) as evaluation metrics. Hits@N denotes the proportion of correct entities in the top N ranks, while MRR is the average reciprocal rank. For both metrics, higher values indicate better alignment results.

4.1.2 Baselines. To verify the effectiveness of IBMEA, we select several typical and competitive methods as baselines. We manifest these methods into two groups: **Traditional Entity Alignment methods.** We choose 6 prominent EA methods proposed in recent years, which achieve entity alignment based on the graph structures and do not introduce multi-modal information, including TransE [7], IPTransE [54], GCN-align [45], KECG [18], BootEA [33], and NAEA [55]. **MMEA methods.** We further collect 10 state-of-the-art MMEA methods, which incorporate entity images as input features to enrich entity representations, including POE [24], Chen

⁴Our **Appendix** and **code** are available at <https://github.com/sutaoyu/IBMEA>.

Table 1: Experimental results on 2 cross-KG datasets where X% represents the percentage of seed alignments used for training. The best result is bold-faced and the runner-up is underlined. * indicates results reproduced using the official source code.

Methods	FB15K-DB15K (20%)			FB15K-DB15K (50%)			FB15K-DB15K (80%)			FB15K-YAGO15K (20%)			FB15K-YAGO15K (50%)			FB15K-YAGO15K (80%)		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
TransE [7]	.078	.240	.134	.230	.446	.306	.426	.659	.507	.064	.203	.112	.197	.382	.262	.392	.595	.463
IPTransE [54]	.065	.215	.094	.210	.421	.283	.403	.627	.469	.047	.169	.084	.201	.369	.248	.401	.602	.458
GCN-align [45]	.053	.174	.087	.226	.435	.293	.414	.635	.472	.081	.235	.153	.235	.424	.294	.406	.643	.477
KECG* [18]	.128	.340	.200	.167	.416	.251	.235	.532	.336	.094	.274	.154	.167	.381	.241	.241	.501	.329
POE [24]	.126	.151	.170	.464	.658	.533	.666	.820	.721	.113	.229	.154	.347	.536	.414	.573	.746	.635
Chen et al. [10]	.265	.541	.357	.417	.703	.512	.590	.869	.685	.234	.480	.317	.403	.645	.486	.598	.839	.682
HMEA [15]	.127	.369	-	.262	.581	-	.417	.786	-	.105	.313	-	.265	.581	-	.433	.801	-
EVA [23]	.134	.338	.201	.223	.471	.307	.370	.585	.444	.098	.276	.158	.240	.477	.321	.394	.613	.471
MSNEA [11]	.114	.296	.175	.288	.590	.388	.518	.779	.613	.103	.249	.153	.320	.589	.413	.531	.778	.620
ACK-MMEA [19]	.304	.549	.387	.560	.736	.624	.682	.874	.752	.289	.496	.360	.535	.699	.593	.676	.864	.744
UMAEA* [13]	.560	.719	.617	<u>.701</u>	.801	.736	<u>.789</u>	.866	.817	<u>.486</u>	.642	<u>.540</u>	.600	.726	.644	.695	.798	.732
MCLEA [22]	.445	.705	.534	.573	.800	.652	.730	.883	.784	.388	.641	.474	.543	.759	.616	.653	.835	.715
MEAformer [3]	<u>.578</u>	<u>.812</u>	<u>.661</u>	.690	<u>.871</u>	<u>.755</u>	.784	<u>.921</u>	<u>.834</u>	<u>.444</u>	<u>.692</u>	.529	<u>.612</u>	<u>.808</u>	<u>.682</u>	<u>.724</u>	<u>.880</u>	<u>.783</u>
IBMEA (Ours)	.631	.813	.697	.742	.880	.793	.821	.922	.859	.521	.708	.584	.655	.821	.714	.751	.890	.800

et al. [10], HMEA [15], EVA [23], ACK-MMEA [19], MSNEA [11], PSNEA [28], UMAEA [13], MCLEA [22], and MEAformer [3]. For more details, please refer to Sec. 5. Among the methods, MCLEA and MEAformer are typical competitive methods.

4.1.3 Implementation Details. In our experiments, the GAT has two layers with a hidden size of $d_g = 300$. For visual embeddings, following [3, 22, 23], we leverage a pre-trained VGG-16 model [31] for cross-KG datasets with $d_v = 4096$ and ResNet-152 for bilingual datasets with $d_v = 2048$ to obtain initial features from the penultimate layer. Following [3, 22], the Bag-of-Words method is chosen to encode both attributes ($\mathbf{x}_a$) and relations ($\mathbf{x}_r$) as fixed-length vectors, where d_a and d_r are both 1000. The graph embedding output is 300, and other modality embeddings are 100. Training is conducted over 1000 epochs with a batch size of 7,500, using AdamW optimizer [26] with a learning rate of $6e-3$ and a weight decay of $1e-2$. Hyper-parameters $\beta_g, \beta_v, \beta_a, \beta_r$ in Eq. (7) are tune in $[1e-4, 1e-3, 1e-2, 1e-1]$, yielding optimal results at $1e-3, 1e-2, 1e-2, 1e-2$, respectively. Consistent with prior studies [3, 22, 23], we adopt an iterative training strategy to overcome the lack of training data and exclude entity names for fair comparison. Our best hyper-parameters are tuned by grid search according to the prediction accuracy of MMEA, detailed in **Appendix**.

4.2 Overall Results

The overall average results on cross-KG and bilingual datasets are displayed in Table 1 and Table 2, respectively. From the tables, we have several observations: 1) *Our proposed IBMEA model outperforms all baseline models on 5 benchmarks with different data settings*, including two cross-KG datasets and three bilingual datasets, excelling in all key metrics (H@1, H@10, and MRR), which demonstrates the model’s generality. Specifically, under 50% and 80% alignment seed settings on 2 cross-KG datasets, our model achieved an average increase of 4.2% and 3.0% in H@1 scores, and 3.5% and 2.1% in MRR scores, respectively. Moreover, our model still exceeds those high-performing baselines and increases the current SOTA Hits@1 scores from .847/.842/.845 to .859/.856/.864 on ZH-EN/JA-EN/FR-EN datasets with the DBP15K, respectively. 2) *Our model achieves better results in relatively low-resource data scenario*. Compared to the

Table 2: Experimental results on 3 bilingual datasets . The best result is bold-faced and the runner-up is underlined. * indicates results reproduced using the official source code.

Methods	DBP15K _{ZH-EN}			DBP15K _{JA-EN}			DBP15K _{FR-EN}		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
GCN-align [45]	.434	.762	.550	.427	.762	.540	.411	.772	.530
KECG [18]	.478	.835	.598	.490	.844	.610	.486	.851	.610
BootEA [33]	.629	.847	.703	.622	.854	.701	.653	.874	.731
NAEA [55]	.650	.867	.720	.641	.873	.718	.673	.894	.752
EVA [23]	.761	.907	.814	.762	.913	.817	.793	.942	.847
MSNEA [11]	.643	.865	.719	.572	.832	.660	.584	.841	.671
UMAEA* [13]	.807	.967	.867	.806	.967	.832	.820	.979	.881
PSNEA [28]	.816	.957	.869	.819	.963	.868	.844	<u>.982</u>	.891
MCLEA [22]	.816	.948	.865	.812	.952	.865	.834	.975	.885
MEAformer [3]	<u>.847</u>	<u>.970</u>	<u>.892</u>	<u>.842</u>	<u>.974</u>	<u>.892</u>	<u>.845</u>	.976	<u>.894</u>
IBMEA (Ours)	.859	.975	.903	.856	.978	.902	.864	.985	.911

runner-up method results, our model achieves an average increase of 4.4% in H@1 and 4.0% in MRR on cross-KG datasets with a 20% alignment seed setting. It demonstrates that using the IB principle, the model can better grasp alignment-relevant information, alleviating overfitting on pieces of predictive clues with limited seed alignments. 3) *The MMEA baseline model generally outperforms the EA baseline*. Remarkably, our model utilizing 20% of the seeds surpasses the H@1 performance of the EA baseline at 80% in cross-KG datasets. It demonstrates that introducing multi-modal information can enrich entity information, and significantly help entity alignment with limited seed alignments. All the results demonstrate the effectiveness of our proposed IBMEA, and we will further examine our model on hard data scenarios to show the results in Sec. 4.4.

4.3 Ablation Study

To evaluate the unity of all regularizers, we conduct ablation study on FB15K-DB15K dataset. From the Table 3, we can observe that: 1) The removal of any modal-specific IB regularizer consistently leads to significant reductions across all metrics, thereby validating its efficacy for modal-specific IB regularizers. 2) As the proportion of alignment seeds increases, the negative impact of removing the

Table 3: Ablation study on IB regularizers. G-IB, V-IB, A-IB, and R-IB are modal-specific IB regularizers (Graph, Visual, Attribute, and Relation) for short. Hybrid-IB refers to using an IB regularizer on the representations after fusion.

Model	FB15K-DB15K (20%)			FB15K-DB15K (50%)			FB15K-DB15K (80%)		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
IBMEA	.631	.813	.697	.742	.880	.793	.821	.922	.859
w/o G-IB	.597	.780	.660	.707	.852	.759	.806	.903	.843
w/o V-IB	.623	.785	.680	.725	.862	.774	.816	.908	.850
w/o A-IB	.613	.782	.672	.721	.858	.771	.807	.898	.840
w/o R-IB	.625	.799	.686	.720	.862	.772	.816	.914	.851
Hybrid-IB	.468	.614	.518	.643	.762	.686	.748	.837	.780

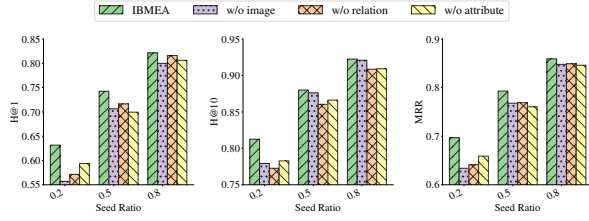


Figure 3: Results of removing different modalities on FB15K-DB15K dataset. w/o means removing the modality.

regularizer diminishes. We believe there may exist more misleading information in low-resource settings, thus the modal-specific IB regulariser has a greater effect. 3) Compared with other w/o variants, w/o G-IB variant exhibits the most dramatic decline in overall results, indicating that the graph structure information plays an important role in MMEA. 4) Comparing hybrid-IB and IBMEA, we see that hybrid-IB results are much lower than IBMEA results, which indicates applying the IB to each modality before fusion is more effective than post-fusion. The IB regularizer used after fusion might indirectly refine essential multi-modal information, thereby hindering the model’s capacity to effectively retain alignment-relevant information while suppressing alignment-irrelevant information.

To further evaluate the impact of different modalities, we report the results by deleting different modalities on FB15K-DB15K dataset, shown in Figure 3. We notice that: 1) Removing each modality will reduce the final result, proving each modality’s importance. 2) Compared with the high proportion of seeds (50% and 80% ratio), the effect of different modalities information on the overall performance is more obvious in the low proportion of seeds. This indicates that our model can extract crucial information of each modality, thus obtaining better results in the case of the low proportion of seeds.

4.4 Further Analysis

4.4.1 Performance on extreme low-resource training data. To further explore the performance with few training data, where the alignment seed ratio ranges from 5% to 30%. Specifically, we select MCLEA and MEAformer as baselines, and the results are depicted in Figure 4. We can observe that as the proportion of alignment seeds decreases, the performance of all models would also decrease

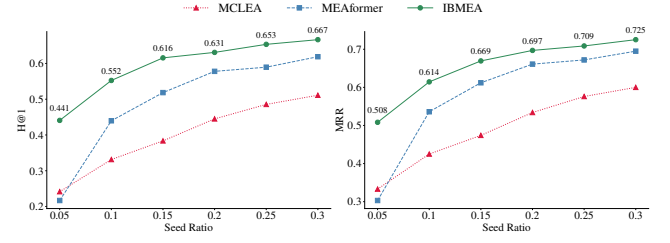


Figure 4: Results in the low-resource data scenario with proportions of seed alignments on FB15K-DB15K dataset.

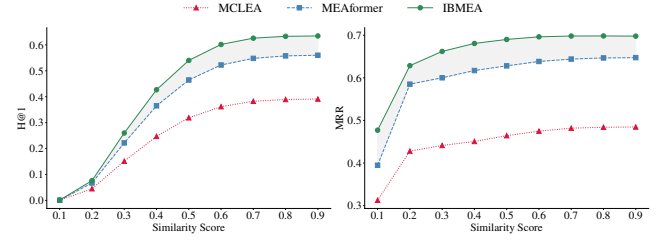


Figure 5: Results on the samples with low-similarity image in FB15K-DB15K dataset.

in terms of metrics. However, it is obvious that our IBMEA continuously outperforms MCLEA and MEAformer, indicating the effectiveness of our proposed method. Moreover, it is worth noting that the gap between them is much more significant when the seed alignments are extremely few (5%), which confirms the superiority of our proposed method to alleviate the overfitting of shortcut misleading clues in low-resource scenarios.

4.4.2 Performance on samples with low-similarity image. To examine the effectiveness of our models on hard samples, we select samples with low-similarity image in FB15K-DB15K dataset (20% seed alignments) and examine the performance of the three MMEA models on these samples. The similarity between the images is calculated by the cosine similarity of the features extracted from the pre-trained visual encoder. Considering alignment entity pairs with low-similarity image, they tend to contain more task-irrelevant misleading information. As shown in Figure 5, our model achieves the best results on both the H@1 and MRR metrics with different entity image similarity settings. We believe that our model can effectively leverage IB principle in training, which encourages the model to focus more on the task-relevant image information for prediction rather than the overall redundant image information.

4.4.3 Performance on samples with high-noise image. To explore the efficacy of entity alignment in noisy data scenarios, we utilize FB15K-DB15K dataset (20% seed alignments) and add artificial noise to the images for experiments. Specifically, we first obtain the entity’s image embedding from the pre-trained visual encoder, and then impose random dropout to them with a dropout rate to control the noise rate. As depicted in Figure 6, results indicate that all models reflect a performance decline with increasing the image

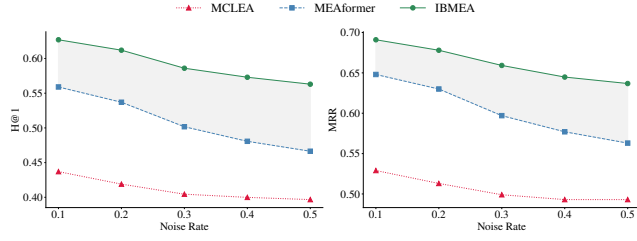


Figure 6: Results on the samples with noisy image in FB15K-DB15K dataset. We impose random dropout to images as noise and control the dropout rate for analysis.

noise rate. However, even under a higher noise rate, our model maintains relatively high H@1 and MRR metrics. Considering the higher noise rate tends to have more misleading clues, the results demonstrate the effectiveness of our model with IB regularizers, which successfully preserves alignment-relevant information while suppressing alignment-irrelevant information, thereby sustaining superior alignment performance in noisy environments.

4.4.4 Case study. To detail the model prediction, we present typical cases of aligned entities from DBP15K_{ZH-EN} with relatively low-similarity figures in Figure 7. Case 1 shows characters with different clothes but similar facial features. Case 2 displays maps with different background colors yet identical contour features. In Case 3, a statue and a black-and-white photo of the same person exhibit similar facial traits, providing crucial information for entity alignment. The three cases reflect misleading clues in images. As in the predictions, it was found that MCLEA and MEAformer predict the true entity with a lower ranking. Note that in Case 3, the MEAformer model predicts worse, likely due to being misled by differences in entity forms. In contrast, our model accurately predicted in all cases, thus demonstrating its superior ability to handle misleading information for the MMEA task.

5 Related work

5.1 Entity Alignment

Entity Alignment (EA) aims to identify equivalent entities across different knowledge graphs (KGs) to facilitate knowledge fusion. The majority of existing EA methods focus on traditional knowledge graphs can be classified into two categories: 1) *Structure-based* techniques focus on employing knowledge embedding [1, 7, 12, 33, 54] to capture the entity structure information from relational triples, or utilize graph-based models [18, 27, 34, 45] for neighborhood entity feature aggregation [17, 21, 30, 40]. 2) *Side information-based* methods [2, 25, 32, 37, 46] integrate side information (e.g., entity name, entity attribute, relation predicates) to learn more informative entity representation. As previous studies [25, 51] have shown, the *structure-based* techniques assume that aligned entities should share similar neighborhoods, the *side information-based* methods regard equivalent entities usually contain similar side information. Nevertheless, these methods mainly solely on structural information and utilize textual side information, making them incapable of handling the visual data of entities in actual scenes.

Case Entity	MMKG-1 Image	MMKG-2 Image	ImgSim	Rank in Prediction		
				MCLEA	MEAformer	IBMEA
① Esteban Granero			0.76	14	13	1
② List of counties in Oregon			0.74	9	13	1
③ Run Run Shaw			0.57	12	46	1

Figure 7: Case study on entities with image similarity.

5.2 Multi-Modal Entity Alignment

Due to the increasing popularity of multi-modal knowledge graphs, how to incorporate visual modalities in EA, i.e. multi-modal entity alignment (MMEA), has attracted research attention. The pioneer method PoE [24] defines overall probability distribution as the product of all uni-modal experts. Chen et al. [10] design a multi-modal fusion module for integrating different modal embeddings. HMEA [15] combines the structure and visual representations in hyperbolic space. MSNEA [11] and XGEA [47] integrate visual features to guide relational and attribute learning, with the former using a translation-based KG embedding method and the latter using a graph neural network (GNN) approach. ACK-MMEA [19] proposes a method for uniformizing multi-modal attributes, while UMAEA [13] introduces multi-scale modality hybrid and circularly missing modality imagination. PSNEA [28] and PCMEA [42] dynamically generate pseudo-label data to improve alignment performance. EVA [23] allows the alignment model to obtain the importance of different modality weights to fuse multi-modal information from KGs into a joint embedding. MCLEA [22] follows the EVA fusion method and then obtains informative entity representations based on contrastive learning. MEAformer [3] develops UMAEA and dynamically learns modality weights for each entity via the transformer-based attention fusion method. However, the above existing methods rarely explicitly address the misleading alignment-irrelevant information in each modality, resulting in incomplete utilization of the alignment-relevant information.

6 Conclusion

In this paper, we explore MMEA to identify equivalent entities between MMKGs. To address alignment-irrelevant misleading clues in modalities, we propose a novel MMEA framework termed as IBMEA, which aims to emphasize alignment-relevant information and simultaneously suppress alignment-irrelevant information in modalities. Particularly, we devise multi-modal variational encoders to represent multi-modal information with probability distributions and propose information bottleneck regularizers and an information contrastive regularizer. Experimental results indicate that our model outperforms previous state-of-the-art methods, and shows promising capability for misleading information alleviation, especially in the low-resource and high-noise data scenarios.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No.2021YFB3100600), the Youth Innovation Promotion Association of CAS (No.2021153), and the Postdoctoral Fellowship Program of CPSF (No.GZC20232968).

References

- [1] 2019. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *Proceedings of ISWC*. Springer, 612–629.
- [2] 2020. COTSAE: co-training of structure and attribute embeddings for entity alignment. In *Proceedings of AAAI*, Vol. 34. 3025–3032.
- [3] 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the ACM MM*. 3317–3327.
- [4] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. 2018. Fixing a broken ELBO. In *Proceedings of ICML*. PMLR, 159–168.
- [5] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *Proceedings of ICLR*.
- [6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual Information Neural Estimation. In *Proceedings of ICML*.
- [7] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of NeurIPS*. 2787–2795.
- [8] Malik Boudiaf, Jérôme Rony, Intiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. 2020. A Unifying Mutual Information View of Metric Learning: Cross-Entropy vs. Pairwise Losses. In *Proceedings of ECCV*, Vol. 12351. 548–564.
- [9] Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. 2022. Cross-modal Knowledge Graph Contrastive Learning for Machine Learning Method Recommendation. In *Proceedings of the ACM MM*. 3694–3702.
- [10] Liyi Chen, Zhi Li, Yijun Wang, Tong Xu, Zhefeng Wang, and Enhong Chen. 2020. MMEA: entity alignment for multi-modal knowledge graph. In *Proceedings of KSEM*. Springer, 134–147.
- [11] Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022. Multi-modal siamese network for entity alignment. In *Proceedings of KDD*. 118–126.
- [12] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment. In *Proceedings of IJCAI*. 1511–1517.
- [13] Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z. Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023. Rethinking Uncertainty Missing and Ambiguous Visual Modality in Multi-Modal Entity Alignment. In *Proceedings of ISWC*. 121–139.
- [14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 11 (2020), 665–673.
- [15] Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing* 461 (2021), 598–607.
- [16] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of ICLR*.
- [17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv:1609.02907* [cs.LG]
- [18] Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised Entity Alignment via Joint Knowledge Embedding Model and Cross-graph Model. In *Proceedings of EMNLP*. 2723–2732.
- [19] Qian Li, Shu Guo, Yangyifei Luo, Cheng Ji, Lihong Wang, Jiawei Sheng, and Jianxin Li. 2023. Attribute-Consistent Knowledge Graph Representation Learning for Multi-Modal Entity Alignment. In *Proceedings of WWW*. 2499–2508.
- [20] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. 2024. A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
- [21] Ke Liang, Lingyuan Meng, Sihang Zhou, Wenxuan Tu, Siwei Wang, Yue Liu, Meng Liu, Long Zhao, Xiangjun Dong, and Xinwang Liu. 2024. MINES: Message Intercommunication for Inductive Relation Reasoning over Neighbor-Enhanced Subgraphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10645–10653.
- [22] Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal Contrastive Representation Learning for Entity Alignment. In *Proceedings of COLING*. 2572–2584.
- [23] Fangyu Liu, Muhao Chen, Dan Roth, and Nigel Collier. 2021. Visual pivoting for (unsupervised) entity alignment. In *Proceedings of AAAI*, Vol. 35. 4257–4266.
- [24] Ye Liu, Hui Li, Alberto García-Durán, Mathias Niepert, Daniel Oñoro-Rubio, and David S. Rosenblum. 2019. MMKG: Multi-modal Knowledge Graphs. In *Proceedings of ESWC*, Vol. 11503. 459–474.
- [25] Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. 2020. Exploring and Evaluating Attributes, Values, and Structures for Entity Alignment. In *Proceedings of EMNLP*. 6355–6364.
- [26] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.
- [27] Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. MRAEA: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *Proceedings of WSDM*. 420–428.
- [28] Wenxin Ni, Qianqian Xu, Yangbangyan Jiang, Zongsheng Cao, Xiaochun Cao, and Qingming Huang. 2023. PSNEA: Pseudo-Siamese Network for Entity Alignment between Multi-modal Knowledge Graphs. In *Proceedings of the ACM MM*. 3489–3497.
- [29] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *Proceedings of ICML*. PMLR, 5171–5180.
- [30] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Proceedings of ESWC (Lecture Notes in Computer Science*, Vol. 10843). 593–607.
- [31] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of ICLR*, Yoshua Bengio and Yann LeCun (Eds.).
- [32] Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-Lingual Entity Alignment via Joint Attribute-Preserving Embedding. In *Proceedings of ISWC*. 628–644.
- [33] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding. In *Proceedings of IJCAI*. 4396–4402.
- [34] Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of AAAI*, Vol. 34. 222–229.
- [35] Naftali Tishby, Fernando C Pereira, and William Bialek. 1999. The information bottleneck method. *Annual Allerton Conf. on Communication, Control, and Computing* (1999).
- [36] Naftali Tishby and Noga Zaslavsky. 2015. Deep Learning and The Information Bottleneck Principle. In *IEEE Information Theory Workshop (ITW)*.
- [37] Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of AAAI*, Vol. 33. 297–304.
- [38] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine* 25, 6 (2008), 131–146.
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748* [cs.LG]
- [40] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *Proceedings of ICLR*.
- [41] Petar Velickovic, William Fedus, William L. Hamilton, Pietro Liò, Yoshua Bengio, and R. Devon Hjelm. 2019. Deep Graph Infomax. In *Proceedings of ICLR*.
- [42] Luyao Wang, Pengnian Qi, Xigang Bao, Chunlai Zhou, and Biao Qin. 2024. Pseudo-Label Calibration Semi-supervised Multi-Modal Entity Alignment. In *Proceedings of AAAI*. 9116–9124.
- [43] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning. In *Proceedings of CVPR*. 5022–5030.
- [44] Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan E Kuruoglu, and Yefeng Zheng. 2020. Information Theoretic Counterfactual Learning from Missing-Not-At-Random Feedback. In *Proceedings of NeurIPS*.
- [45] Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of EMNLP*. 349–357.
- [46] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2019. Jointly Learning Entity and Relation Representations for Entity Alignment. In *Proceedings of EMNLP*. 240–249.
- [47] Baogui Xu, Chengjin Xu, and Bing Su. 2023. Cross-Modal Graph Attention Network for Entity Alignment. In *Proceedings of the ACM MM*. 3715–3723.
- [48] Guohai Xu, Hehong Chen, Feng-Lin Li, Fu Sun, Yunzhou Shi, Zhixiong Zeng, Wei Zhou, Zhongzhou Zhao, and Ji Zhang. 2021. Alime mkg: A multi-modal knowledge graph for live-streaming e-commerce. In *Proceedings of CIKM*. 4808–4812.
- [49] Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning Cross-Lingual Entities with Multi-Aspect Information. In *Proceedings of EMNLP*. 4430–4440.
- [50] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the ACM MM*. 1942–1951.
- [51] Rui Zhang, Bayu Distiawan Trisedya, Miao Li, Yong Jiang, and Jianzhong Qi. 2022. A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning. *The VLDB Journal* 31, 5 (2022), 1143–1168.

- [52] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Multi-modal Knowledge-aware Hierarchical Attention Network for Explainable Medical Question Answering. In Proceedings of the ACM MM. 1089–1097.
- [53] Fei Zhao, Chunhui Li, Zhen Wu, Shangyu Xing, and Xinyu Dai. 2022. Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER. In Proceedings of the ACM MM. 3983–3992.
- [54] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative Entity Alignment via Joint Knowledge Embeddings. In Proceedings of IJCAI. 4258–4264.
- [55] Qiannan Zhu, Xiaofei Zhou, Jia Wu, Jianlong Tan, and Li Guo. 2019. Neighborhood-Aware Attentional Representation for Multilingual Knowledge Graphs. In Proceedings of IJCAI. 1943–1949.